ELSEVIER

FEBS Letters 339 (1994) 269–275

FEBS 13677

FEBS LETTERS

A mutation data matrix for transmembrane proteins

D.T. Jones[a,b,*], W.R. Taylor[b], J.M. Thornton[a]

[a] *Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK*

[b] *Laboratory of Mathematical Biology, The National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*

Received 23 December 1993

Abstract

The widely used Mutation Data Matrix (MDM), is an amino acid comparison matrix calculated from a study of the exchange probabilities (or odds) derived from an analysis of the evolutionary changes seen in groups of very similar proteins. In this work, a mutation data matrix is calculated for membrane spanning segments. This new mutation data matrix is found to be very different from matrices calculated from general sequence sets which are biased towards water-soluble globular proteins, and the differences are discussed in the context of specific structural requirements of membrane spanning segments. This new matrix will help improve the accuracy of integral membrane protein sequence alignments, and could also be of use in the rational design of site directed mutagenesis experiments for this class of proteins

Key words: Amino acid; Comparison matrix; Mutation; Evolution; Membrane protein; Transmembrane segment

1. Introduction

Given the extreme difference between the typical environment of integral membrane associated proteins and that of globular proteins, it is not surprising that the relationship between protein sequence and structure is different for these two important classes of proteins. A very obvious example of this is the difference in the structural roles played by the 20 standard amino acids in transmembrane segments and in globular domains. A simple way to analyse amino acid properties is to observe the frequencies of amino acid exchanges in closely related sequences, a technique typified by the ubiquitous Dayhoff matrix calculated by Dayhoff et al. [1], which is widely used in sequence comparison applications. We have recently described a highly efficient method for generating mutation data matrices from very large sequence sets [2], and have applied this method to the generation of a matrix based on mutations occurring in transmembrane segments of integral membrane proteins.

* Corresponding author. Fax: (44) (71) 380 7193.

Copies of the complete data, including all intermediate matrices required for constructing matrices other than the 250 PAM matrix and matrices for single and multi-spanning segments, may be obtained from the authors in printed or machine readable form.

2. Materials and methods

Our method for generating a mutation data matrix is very similar in essence to that described by Dayhoff et al. [1]. The method involves three steps: (a) clustering the sequences into homologous families, (b) tallying the observed mutations between highly similar sequences, and (c) relating the observed mutation frequencies to those expected by pure chance. The main difference here is in our use of an approximate method (a pairwise present-day ancestor scheme) for inferring the phylogenetic relationships amongst the sequences in the data set. A program was written to compute all the relevant data automatically from a file of protein sequences [2].

Before tree construction can begin, it is necessary to generate a similarity matrix. Evidently, since only very closely related proteins are used in the derivation, the vast majority of pairwise comparisons are unnecessary, so some simple (and quick) means is needed to filter out sequence pairs that have no chance of producing alignment scores $\geq 85\%$ identity. A simple approximate algorithm is used for 'estimating' the percentage identity between two protein sequences without prior alignment [2]. The algorithm considers the distribution of amino acid triplets (or 3-tuples) between the two sequences. If there are sufficient identical triplets between both sequences we assume that the sequences show a potential homology. Taking the longest sequence, a hash table is constructed containing the frequencies of occurrence of the constituent triplets, and using this table, the triplet frequencies of the shorter sequence are then compared with those of the longer and a comparison score calculated.

By aligning only those sequence pairs with corrected triplet scores indicating sequence identity $\geq 45\%$ and subsequently excluding sequence pairs with rigorous alignment scores of $\leq 85\%$ identity we were able to rapidly cluster the sequences. By combining this very rapid heuristic measure of identity with an efficiently coded dynamic programming algorithm [3] as a 'second level filter' we were able to construct the similarity matrix at an average rate of over 1000 similarity score calculations per second on a desktop workstation.

Three possible methods were considered for selecting a suitable transmembrane data set. Ideally, the data set would be constructed from all the segments *experimentally determined* to be transmembranal (either where the 3D structure is known or where the membrane topology studies has been studied by chemical or immunological means). Failing that ideal, putative transmembrane segments could be included, i.e.

SSDI 0014-5793(94)00061-Y

segments which have been identified as probably transmembranal by the sequence depositors, either through a knowledge of the relevant biochemistry, by homology or analogy with a related protein, or through standard prediction techniques [4,5]. The third option would be to extend the data set further by applying a standard prediction algorithm to undocumented sequences expected to include transmembrane segments. The first approach is at present not feasible due to the very limited experimental data on integral membrane proteins Despite the success of current prediction techniques [5,6] they are still not reliable enough to apply blindly, and therefore we rejected this option in favour of using documented transmembrane segments, including those which are experimentally determined and those which have been essentially predicted, but which have at least been vetted by the sequence depositors.

The source data for this work was a set of documented transmembrane segments extracted from Release 23.0 of SWISS-PROT [7]. This derived databank comprised 1765 sequences, containing 5662 transmembrane segments. This dataset was extended by searching for sequences closely related (≥ 85% sequence identity) to this initial set in a minimally redundant sequence databank (D.T. Jones, unpublished results). This databank comprises all the non-identical protein sequences extracted from SWISS-PROT Release 23, PIR Release 33 [7] and an automatic translation of GenBank Release 73 [8], totalling 72,000 sequences. Using the MAKEPET program [2], a mutation data analysis was performed on this final data set, and a set of mutation data matrices calculated. The final matrix was generated from 3155 pairwise alignments (in total, 1.27×10^8 sequence comparisons were performed), providing 4845 accepted point mutations (PAMs). Separate analyses were performed for both single-spanning (1765 alignments, 1765 PAMs) and multiple-spanning transmembrane segments (1405 alignments, 3612 PAMs). The combined transmembranal matrix is based on 3 times as many PAMs as the Dayhoff matrix, but in view of the fact that some amino acids occur very infrequently in transmembrane segments, such a large data set is essential to provide sufficient samples across the entire matrix.

3. Results and discussion

The previously observed amino acid biases in transmembrane segments [9] are evident in Table 1. The most commonly occurring residue in transmembrane helices is leucine both for single and multi-spanning segments. Valine in the next most common residue in single-spanning segments, and isoleucine the next most common residue in multi-spanning segments. As expected, the polar residues are not frequent in transmembrane segments, with the negatively charged amino acids being the most clearly disfavoured residues. Single-spanning segments are significantly more hydrophobic in nature than multi-spanning segments with a total frequency of occurrence of hydrophobic amino acids (alanine, isoleucine, leucine, methionine, phenylalanine, tryptophan, valine) of 68% compared with the multi-spanning frequency of 55%.

The upper half of Table 2 shows how many of each of the possible 190 exchanges were observed in all the transmembrane segments, with the lower half of Table 2

Table 1
Relative mutabilities and normalized frequencies of occurrence for the 20 amino acid residues, calculated from transmembrane protein segments compared with the values calculated from a general set of proteins [2]

	Relative Mutability (General)	Frequency of Occurrence (General)	Relative Mutability (Transmem)	Frequency of Occurrence (Transmem)	Relative Mutability (Single)	Frequency of Occurrence (Single)	Relative Mutability (Multi)	Frequency of Occurrence (Multi)
Ala (A)	100.0	0 0767	100 0	0 1051	100 0	0 1137	100.0	0 1026
Arg (R)	82 7	0 0515	134.3	0 0157	182 1	0 0217	106 8	0 0135
Asn (N)	103 0	0 0427	60 2	0 0185	115 8	0 0109	50 4	0 0211
Asp (D)	84 1	0 0518	76 3	0 0089	56 9	0 0057	79 9	0 0100
Cys (C)	46 2	0 0196	98 7	0 0219	71 8	0 0193	106 5	0 0229
Gln (Q)	84 5	0 0405	80 2	0 0141	121 7	0 0066	74 6	0 0168
Glu (E)	76 3	0 0617	72 3	0 0097	163 9	0 0047	57 0	0 0113
Gly (G)	52 0	0 0733	50 7	0 0758	49 6	0 0888	51 3	0 0712
His (H)	91 9	0 0228	63 9	0 0168	79 3	0 0113	61 2	0 0188
Ile (I)	102 5	0 0539	135 4	0 1188	127 0	0 1326	138 2	0 1137
Leu (L)	54 0	0 0919	69 2	0 1635	58 3	0 1769	72 7	0 1583
Lys (K)	72.5	0 0588	79 7	0 0112	129 2	0 0120	62 6	0 0111
Met (M)	95 6	0 0239	146 3	0 0333	193 7	0 0284	132 6	0 0351
Phe (F)	51 1	0 0402	65.7	0 0777	89 6	0 0554	59 9	0 0856
Pro (P)	58 4	0 0508	42.4	0 0260	84 9	0 0173	35 0	0 0291
Ser (S)	116 4	0 0685	110 2	0 0568	99 0	0 0478	113 8	0 0597
Thr (T)	107 1	0 0586	127 9	0 0523	161 1	0 0499	119 2	0 0531
Trp (W)	25 1	0 0143	38.8	0 0223	80 0	0 0168	28 7	0 0242
Tyr (Y)	48 8	0 0322	48.3	0 0324	79 7	0.0235	40 9	0 0353
Val (V)	100 1	0 0661	144 4	0 1195	121 6	0 1565	155 1	0 1065

Relative to Ala which is arbitrarily assigned a mutability of 100

showing the transmembranal counterpart of the widely used MDM78 matrix (250 PAM $\log_{10}$ relatedness-odds matrix). The 250 PAM matrix (PAM = number of accepted point mutations per 100 residues per unit evolutionary time) is shown here for comparison with the most common variant of the original matrix, and it should be born in mind that matrices calculated for evolutionary distances other than 250 PAMs are often found to perform better for some sequence comparisons. Matrices for other distances may be derived from mutation probability matrices obtained by repeated multiplication of the 1 PAM mutation probability matrix, which is shown in Table 3.

As might be expected, the transmembrane protein mutation data matrix is quite different from the matrix calculated from a general sequence set. The most obvious feature of the matrix is the high relative mutability of the hydrophobic residues: isoleucine, methionine, and valine. Interestingly, leucine (the most commonly occurring residue in transmembrane segments) is roughly half as mutable as the other hydrophobic residues, possibly as a result of its high propensity for helix formation (in globular proteins). It is possible that the presence of leucine (and alanine) helps stabilize the helical conformation both prior to, and after, membrane insertion and is thus more highly conserved than the other hydrophobic residues which are found to disfavour helix formation in solution. An alternative explanation for the relative immutability of leucine could be that it is particularly compatible with the aligned helix packing generally observed in transmembrane proteins, a situation perhaps somewhat akin to that of the leucine-zipper motif [10]. However, the fact that the mutability of leucine is just as low for single-spanning segments would seem to point away from this explanation.

The high propensity for tryptophan to exchange with arginine (28 observed exchanges) is rather surprising. However, these exchanges occur in very few protein families (primarily cytochrome c oxidase polypeptide II, and the ATP synthase A chain) and could therefore be rather atypical of transmembrane segments as a whole. On the other hand, it is possible that tryptophan and arginine could participate in similar interactions with the apolar lipid and the polar head groups. In this situation, the polar epsilon nitrogens in both amino acids could interact favourably with the head groups whilst the preceding apolar sections of the respective side chains could interact favourably with the lipid. As a result of the high

Table 2
The 250 PAM transmembrane protein exchange matrix ($\log_{10}$ relatedness odds), based on 4845 accepted point mutations found in 5662 transmembrane segments. Values have been multiplied by 10 and rounded to the nearest integer. The upper half of the matrix shows the actual numbers of exchanges observed

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	21	2	7	13	4	6	160	6	44	43	5	10	21	34	198	202	0	1	292
R	-1	7	0	1	2	21	3	22	21	4	8	53	19	0	1	5	5	28	0	0
N	-1	2	11	14	1	7	0	0	8	4	5	11	3	1	2	32	19	1	1	2
D	0	1	6	12	0	0	12	15	4	1	0	2	1	0	1	0	6	0	1	4
C	0	-1	-1	-3	6	0	0	13	2	4	11	0	1	34	0	48	13	8	23	47
Q	-2	6	3	2	-3	11	16	1	26	1	16	6	3	0	5	7	2	0	0	0
E	0	2	1	8	-3	7	13	21	0	0	0	0	0	0	0	4	2	0	0	7
G	1	0	-2	3	-1	-1	3	6	1	10	0	0	3	4	7	64	12	5	0	53
H	-3	5	3	3	-1	7	2	-3	11	3	2	0	1	0	0	0	4	0	29	2
I	0	-3	-3	-3	-1	-4	-4	-2	-4	2	273	0	161	66	4	22	150	1	4	883
L	-2	-3	-4	-5	-1	-2	-5	-4	-4	1	3	1	153	251	37	43	26	20	6	255
K	-2	9	5	3	-3	6	1	-1	4	-4	-4	12	4	0	0	1	2	0	5	1
M	-1	0	-2	-3	-1	-2	-3	-3	-3	1	1	-1	3	8	0	1	32	1	5	89
F	-2	-4	-4	-6	1	-4	-6	-4	-3	-1	1	-5	0	5	0	32	9	2	54	37
P	0	-3	-2	-2	-4	0	-3	-2	-4	-3	-1	-4	-3	-4	11	9	10	0	1	1
S	2	-1	2	0	1	-1	0	1	-2	-1	-2	-1	-2	-1	-1	3	134	1	22	13
T	1	-1	1	0	0	-2	-1	0	-2	0	-1	-2	0	-2	-1	2	3	1	3	48
W	-4	5	-3	-4	1	0	-3	-2	-1	-3	-2	3	-2	-3	-6	-3	-4	12	2	18
Y	-3	-1	-1	-2	3	0	-5	-5	6	-4	-3	1	-3	2	-5	0	-3	-2	10	2
V	0	-2	-3	-3	0	-4	-2	-1	-4	2	0	-4	1	-1	-3	-1	0	-2	-4	2

probability of subsequent arginine → lysine exchanges, tryptophan also scores highly with lysine in the 250 PAM log odds matrix despite the fact that no direct tryptophan-lysine exchanges were observed in the current data set.

As expected, proline residues appear to be highly conserved in transmembrane segments, presumably due to the special role of proline residues in 'kinking' transmembrane helices, as noted by two groups [11,12]. It should be noted that the frequency of occurrence of proline in transmembrane segments is not much different from its frequency of occurrence in the general sequence set. However, if it is presumed that most of the transmembrane segments are in fact transmembrane helices, and the frequency of occurrence of proline in these segments (2.6%) is compared to the equivalent frequency of 1.9% in globular protein helices, proline appears somewhat more prevalent in transmembrane helices than in globular protein helices. The difference is even more striking when the occurrence of proline-containing helices is considered: only 19% of helices in globular proteins contain one or more proline residues, whereas 50% of the annotated transmembrane segments were found to incorporate this amino acid. These occurrences become 3.5% and 37%, respectively, if the first turn of the helix is excluded from the calculation. Thus proline occurs in the middle of transmembrane helices 10 times as often as it does in the middle of helices in globular domains.

Apart from serine and threonine, the polar residues in general are less mutable in transmembrane protein segments than their counterparts in globular proteins. Serine and threonine are unusual in that they are capable of satisfying the hydrogen bonding capacity of their single hydroxyl groups by interacting with the main chain carbonyl group of residue $i-3$ or $i-4$ in the previous turn of the helix, and are thus compatible with the lipid environment. In terms of their exchanges with their apolar equivalents (leucine and isoleucine), serine prefers to exchange with leucine whereas threonine prefers isoleucine. This is in accordance with the fact that both threonine and isoleucine have centres of asymmetry, and a similar exchange pattern is observed in the general sequence set. It would appear that for multispanning transmembrane segments, polar residues are fairly highly conserved. Polar residues in these transmembrane segments are generally associated with specific functionality, either binding required prosthetic groups, forming ion-channels or perhaps stabilizing the helical bundles by forming ion-pairs. Polar residues, and in particular charged residues, are so infrequently found in single-spanning segments that mutation data for these residues are not statistically significant. The fact that arginine and lysine

Table 3
Mutation Probability Matrix for an evolutionary distance of 1 PAM. Values are scaled by a factor of 10^5. Elements of this matrix give the probability that a residue in column j will mutate to the residue in row i in an evolutionary distance of 1 PAM. Diagonal elements of this matrix represents the probability of residue i = j remaining unchanged

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98950	138	11	81	61	29	64	218	37	38	27	46	31	28	135	360	399	0	3	252
R	21	98590	0	12	9	154	32	30	129	3	5	487	59	0	4	9	10	129	0	0
N	2	0	99368	162	5	51	0	0	49	3	3	101	9	1	8	58	38	5	3	2
D	7	7	78	99200	0	0	128	20	25	1	0	18	3	0	4	0	12	0	3	3
C	13	13	6	0	98964	0	0	18	12	3	7	0	3	45	0	87	26	37	73	41
Q	4	138	39	0	0	99158	171	1	160	1	10	55	9	0	20	13	4	0	0	0
E	6	20	0	139	0	117	99241	29	0	0	0	0	0	0	0	7	4	0	0	6
G	157	145	0	174	61	7	225	99468	6	9	0	0	9	5	28	116	24	23	0	46
H	6	138	45	46	9	190	0	1	99329	3	1	0	3	0	0	0	8	0	92	2
I	43	26	22	12	19	7	0	14	18	98579	172	0	499	88	16	40	296	5	13	763
L	42	53	28	0	52	117	0	0	12	237	99274	9	475	333	147	78	51	92	19	220
K	5	349	62	23	0	44	0	0	0	0	1	99164	12	0	0	2	4	0	16	1
M	10	125	17	12	5	22	0	4	6	140	97	37	98465	11	0	2	63	5	16	77
F	21	0	6	0	160	0	0	5	0	57	158	0	25	99311	0	58	18	9	172	32
P	33	7	11	12	0	37	0	10	0	3	23	0	0	0	99555	16	20	0	3	1
S	194	33	179	0	226	51	43	87	0	19	27	9	3	42	36	98844	265	5	70	11
T	198	33	106	70	61	15	21	16	25	130	16	18	99	12	40	244	98657	5	10	41
W	0	184	6	0	38	0	0	7	0	1	13	0	3	3	0	2	2	99593	6	16
Y	1	0	6	12	108	0	0	0	178	3	4	46	16	72	4	40	6	9	99493	2
V	287	0	11	46	221	0	75	72	12	767	161	9	276	49	4	24	95	83	6	98485

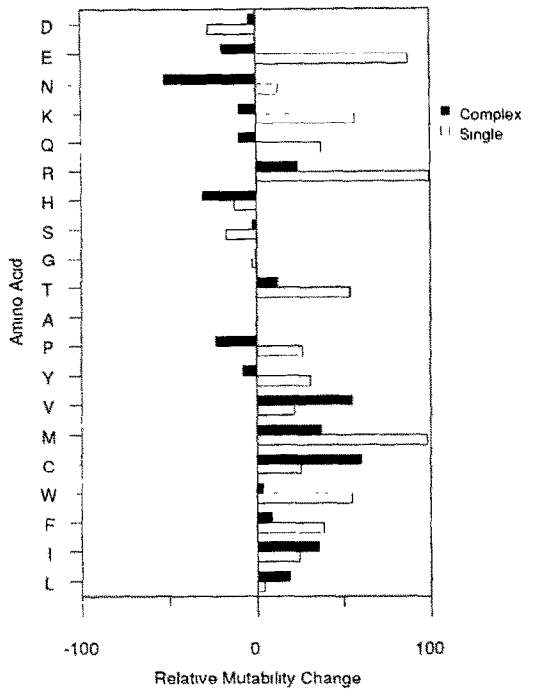

Fig. 1. Changes in relative mutability between general proteins and integral membrane proteins. Data for both single-spanning and complex segments is shown. Positive values indicate that the mutability for transmembrane proteins is higher than that for general sequences. The amino acids are ordered along the y-axis by their polarity [18], with the most polar amino acid at the top.

appear to be fairly mutable might be surprising considering their important role as topogenic signals [5]. However, on closer inspection it is seen that despite being fairly mutable, they tend to exchange between themselves. Presumably, arginine and lysine are equally satisfactory in directing membrane insertion.

General trends in the mutability changes observed in transmembrane segments are clearly seen in Fig. 1. For multi-spanning proteins, a clear distinction is seen between polar and apolar amino acids, where the apolar amino acids become highly variable and the polar amino acids highly conserved. Perhaps the most notable example of this change is the change observed for asparagine, which changes from being one of the third most mutable residues in the general sequence set to being the fourth most highly conserved. A possible reason is again that side chains of asparagine residues are able to hydrogen bond back to their own main chain. In the case of single-spanning segments, there is a much higher background level of mutation than for multi-spanning segments. This is evidenced by the higher average mutation rate for these segments: 0.046 mutations per residue as opposed to 0.029 mutations per residue. Clearly there are far fewer sequence constraints on these segments, and it would appear that the only real requirements for these segments is that they be hydrophobic and contain strong helix-formers.

Despite the fact that the trends in the transmembrane mutation data are as expected from a knowledge of the lipid environment, one of the most important conclusions to be formed from this data is that comparison matrices calculated for general sequence sets do not adequately describe the conservation patterns observed in transmembrane segments. Of course the most important factor in amino acid similarity matrices is the groupings of the side chain chemical properties, which remains constant. However, the relative importance of these properties is seen to be very different for transmembrane segments. These similarities between amino acids are perhaps best visualized by a multi-dimensional projection of the mutation data matrix [13–15]. Fig. 2 shows such a projection of the mutation data matrix in Table 2, and the equivalent matrix for the general sequence set. It is clear from the projections made, that the amino acid groupings are indeed well conserved, yet the separation between groups is somewhat different. In the general sequence set, hydrophobicity and size contribute equally to the conservation patterns observed, whereas size contributes very little to the transmembranal pattern. In the general set, alanine, serine, threonine and proline cluster with the polar residues, whilst in transmembrane segments they are seen to be more closely related to the hydrophobic group. Hydrophobicity is of course by far the most significant factor for transmembrane segments, but the next most important classification to make is whether the side chain is charged, and whether it is negatively charged or positively charged. In the general protein set the charged amino acids cluster together, with little distinction between oppositely charged groups (aspartic acid/lysine for example). In transmembrane segments, however, the sign of the charge is apparently more important, since charged amino acids in these segments are usually functionally-related, or involved in directing the orientation of the segments in the membrane.

The main objective of this study was to investigate the constraints imposed on residue mutation by a lipid envi-

Table 4
Results from searching a set of 2137 transmembrane proteins from SWISS-PROT Release 23 using a dynamic programming algorithm using the sequence of bacteriorhodopsin from halobacterium halobium. Z-scores are given for both the native sequence and sensory rhodopsin I, which is the most distant member of the bacteriorhodopsin family in the data bank

	Standard algorithm	Bipartite Scoring	Helix Gap-penalty	Bipartite Scoring + Helix gap-penalty
Bacteriorhodopsin	28.07	31.46	30.76	33.85
Sensory rhodopsin I	4.61	6.00	5.72	6.07

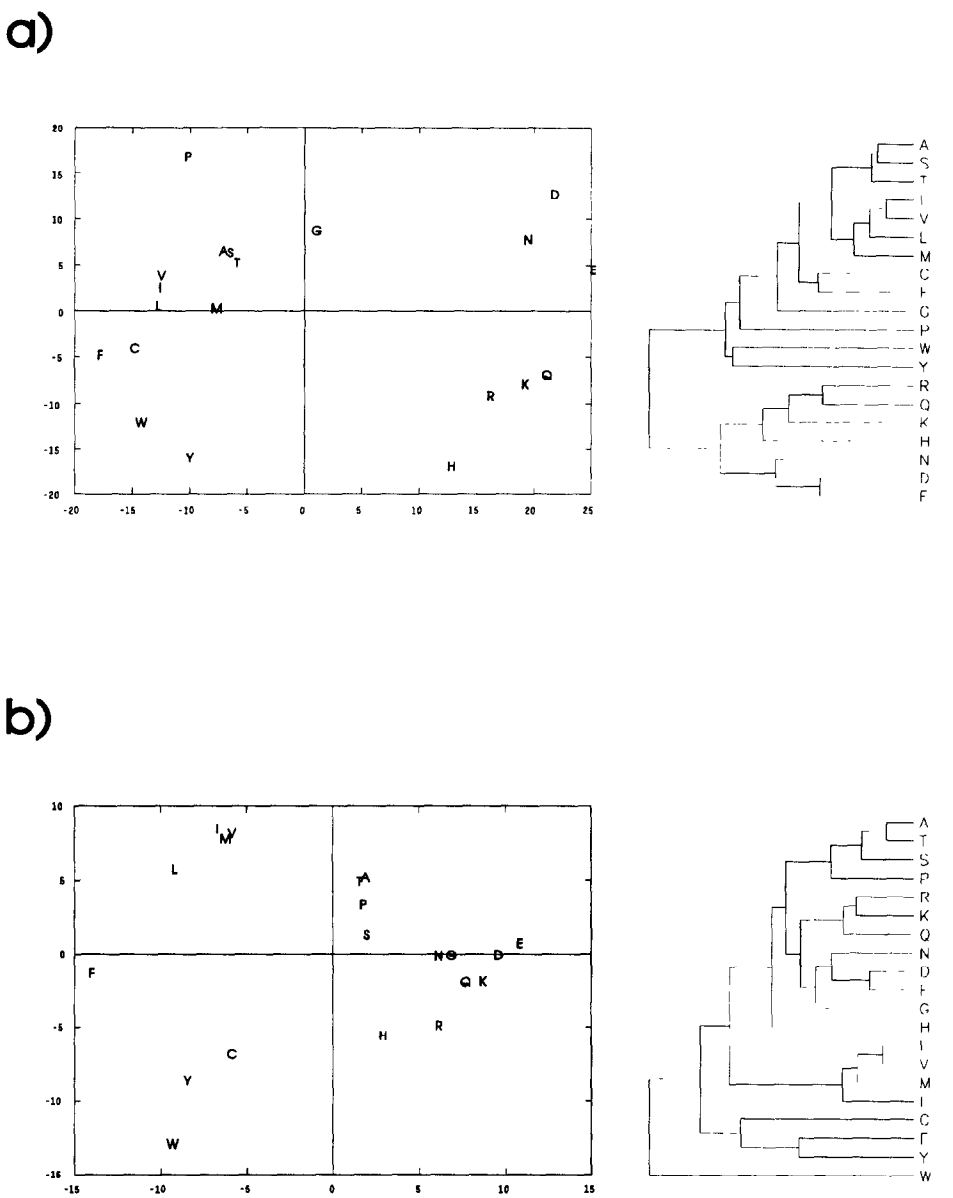

Fig. 2. Multi-dimensional scaling projections of the 250 PAM log-odds matrices, and unweighted pair group mean analysis dendrograms for (a) transmembrane sequences and (b) a general set of sequences [2].

ronment. Despite this it is useful to look at possible applications of the results obtained. Considering the differences in the mutability patterns observed between a lipid and non-lipid environment, clearly when trying to align distantly related transmembrane segments it is vital to bear these differences in mind. Alignment programs that use the transmembrane matrix for transmembrane regions (either experimentally determined or predicted) and a general mutation data matrix for the polar flanking regions are likely to perform much better than programs that use a single matrix. To see if there are any real benefits to be gained from such a bipartite scheme, we applied the following simple test. The sequence of bacteriorhodopsin was taken along with the helix spans as observed in the published three-dimensional structure [16]. Using the dynamic programming algorithm of Gotoh [17] with a constant gap penalty of 15, the bacteriorhodopsin sequence was compared with all the membrane-related protein entries in SWISS-PROT. This search was repeated with a two-matrix scheme, where the matrix given in Table 2 was used for residues aligned with the transmembrane segments of the bacteriorhodopsin, and a general mutation data matrix [2] used for the flanking regions. In addition to this, further searches were performed where the gap-penalty was adjusted depending on whether the gap was in a transmembrane helix. In these cases the gap-penalty of 15 was increased to 150, thus preventing gaps from occurring in transmembrane helices. Table 4 summarizes the results of these searches, where each value represents the Z-score

(number of standard deviations above the mean score) for the alignment between bacteriorhodopsin and the matched sequence. From these results it can be seen that the bipartite scoring scheme alone offers a noticeable improvement in alignment significance, with maximum benefit coming from the additional use of secondary structure specific gap-penalties.

Acknowledgements: We thank Michael Green, Frances Richardson and Dek Woolfson for discussion. This work was supported by an SERC CASE studentship with the MRC, awarded to D.T.J.

References

[1] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) in: Atlas of Protein Sequence and Structure. Vol. 5 suppl. 3, pp. 345–352. National Biomedical Research Foundation, Washington, DC.
[2] Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) Comput. Applic. Biosci. 8, 275–282.
[3] Taylor, W.R. (1988) J. Mol. Evol. 28, 161–169.
[4] Rao, J.K.M. and Argos, P. (1986) Biochim. Biophys. Acta 869, 197–214.
[5] von Heijne, G. (1992) J. Mol. Biol. 225, 487–494.
[6] Jahnig, F. (1990) Trends Biochem. Sci. 15, 93–95.
[7] Bairoch, A. and Boeckmann, B. (1991) Nucleic Acids Res. 19, 2247–2249.
[8] Bilofsky, H.S. and Burks, C. (1988) Nucleic Acids Res. 16, 1861–1863.
[9] von Heijne, G. (1981) Eur. J. Biochem. 120, 275–278.
[10] O'shea, E.K., Klemm, J.D., Kim, P.S. and Alber, T. (1991) Science 254, 539–544.
[11] von Heijne, G. (1991) J. Mol. Biol. 218, 499–503.
[12] Woolfson, D.N., Mortishiresmith, R.J. and Williams, D.H. (1991) Biochem. Biophys. Res. Commun. 175, 733–737.
[13] French, S. and Robson, B. (1983) J. Mol. Evol. 19, 171–175.
[14] Taylor, W.R. (1986) J. Theor. Biol. 119, 205–218.
[15] Taylor, W.R. and Jones, D.T. (1993) J. Theor. Biol., in press.
[16] Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E. and Downing, K.H. (1990) J. Mol. Biol. 213, 899–929.
[17] Gotoh, O. (1982) J. Mol. Biol. 162, 705–708.
[18] Grantham, R. (1974) Science 185, 862–864.